**HOW DO YOU BUILD EFFECTIVE DATA CLASSIFICATION AND HANDLING OF DOCUMENTS**

**LOVETH LUMUMBA ADEH**

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**RSCH6016: CAPSTONE PROJECT**

**PROFESSOR ROSIE NANJI**

**MAY 21, 2023**

**Table of Contents**

**Abstract 3**

**Background of the Study 4**

**Research Question 4**

**Literature Review……………………………………………………………………………………… 5**

**Methodology 10**

**Research Design 10**

**Data Collection 11**

**Data Analysis 11**

**Summary 12**

**References 13**

# **Abstract**

Integrating human expertise and automation technologies is essential for data classification and document handling. This integration enhances accuracy by leveraging human decision-making and automation's computational power. It improves efficiency by automating routine tasks and freeing up human experts for more complex work. Lastly, it strengthens security by combining human understanding of regulations and policies with automated enforcement of security measures.

# **BACKGROUND OF THE STUDY**

The exponential proliferation of data in the digital age has presented organizations with a never-before-seen problem when it comes to protecting sensitive information and upholding conformity to data protection laws. As a result, there is a growing need for reliable and effective data classification and document handling procedures. To minimize data breaches, uphold confidentiality, guarantee data integrity, and comply with legal standards, an organization's information security policy must include effective classification and handling of documents.

Data classification is the process of organizing data into categories for its most effective and efficient use. It provides an interface for organizations to implement controls and procedures across data formats, structures, and storage technologies. Classified data allows an organization to define and implement a single policy for handling sensitive data across multiple systems and data objects. (Satori Cyber, 2023)

For the security and privacy of information in diverse domains and situations, data classification and document management are critical activities. Data classification, per Microsoft Service Assurance, is "the process of identifying, categorizing, and protecting content according to its sensitivity or impact level" (Gluckd & Rubmazz, 2023).

Depending on the value and risk of their data, organizations may apply the proper security controls and policies by classifying the data. The administration of documents from the time they are created until they are destroyed is referred to as document handling. Ilia Sotnikov claims that data classification is the process of classifying data that is structured as well as unstructured into predetermined categories that reflect various forms of data (Sotnikov, 2020).

Public, Confidential, Sensitive, and Personal are among the standard categories used in data categorization, according to Ilia Sotnikov (Sotnikov, 2020). Data that is limited to usage by persons or groups is referred to as sensitive data in general. It's common to use the terms sensitive and confidential data interchangeably. Intellectual property and trade secrets are two examples of sensitive data. To handle documents effectively and securely, it is necessary to store, retrieve, share, archive, and delete them. Particularly in the age of big data, cloud computing, and remote systems, data categorization and document handling provide several issues for organizations (Scarfone et al., 2021).

The complexity, diversity, and volume of data sources and formats, the dynamic and evolving nature of data sensitivity and impact, the adherence to various laws and standards for data protection, and the integration of human expertise and automation technologies for data-centric security management are some of these challenges.

# **AIMS AND OBJECTIVES**

The purpose of this study is to investigate the most effective techniques and best practices for integrating human experience with automated technology in data categorization and document processing to achieve maximum accuracy, efficiency, and security.

The specific objectives are:

●      To review the existing literature on data classification and document handling methods and models.

●      To identify the key factors and criteria that influence the integration of human expertise and automation technologies in data-centric security management.

●      To propose a framework for designing and implementing data classification and document handling systems that leverage human expertise and automation technologies.

●      To evaluate the performance and usability of the proposed framework using a case study approach.

# **RESEARCH QUESTION**

What are the most effective strategies and best practices for integrating human expertise and automation technologies in data classification and document handling to achieve optimal accuracy, efficiency, and security?

LITERATURE REVIEW

# **Data Classification and Handling of Documents**

In the context of a zero-trust strategy, which attempts to secure data regardless of its location or sharing status, data classification and document handling are crucial components of data security and governance. Data handling is the process of implementing the proper rules and controls to data based on its classification, whereas data classification is the process of identifying, categorizing, and labeling data by its sensitivity or impact level. This literature review will explore the benefits, challenges, and best practices of data classification and handling of documents.

# **Benefits of Data Classification and Handling of Documents**

Data classification and handling of documents can provide several benefits for organizations: (1) Lowering the possibility of data breaches, unintentional disclosures, or unauthorized changes, hence improving data security and compliance. Data categorization may assist organizations in identifying and prioritizing their most important and sensitive data so that they can get the best possible security and oversight. Data handling may assist organizations in making sure that data is compliant and securely kept, transported, accessed, and disposed of following its categorization level. Microsoft Service Assurance advises utilizing sensitivity labels to categorize and safeguard data in Microsoft 365 apps, such as encrypting extremely sensitive documents or restricting their transmission via email or printing (Gluckd & Robmazz, 2023). Similarly, to this, Imperva advises utilizing data masking techniques to anonymize private information for third parties or non-production settings (Chugh et al., 2021), (2) Improving data quality and usefulness by eliminating redundancy, inconsistency, and irrelevance. Data categorization may assist organizations in more efficiently organizing and managing their data by removing unneeded or redundant data, standardizing data formats and metadata, and improving data discovery and retrieval. By ensuring that data is backed up, updated, archived, or destroyed as appropriate, data management may assist organizations in maintaining the integrity and availability of their data. According to Gartner, a data lifecycle management framework should be used to identify the phases and activities for various types of data depending on their value and utilization (Scarfone et al., 2021). Varonis also recommends utilizing data retention policies to automate the deletion or archiving of old or obsolete data (Cocchi et al., 2018), and (3) Supporting data sharing and cooperation to support innovation and decision-making based on data. By giving various individuals or groups the proper access rights and permissions based on their roles and responsibilities, the categorization of data enables organizations to strike a balance between data protection and data usefulness. By ensuring that data is interoperable, compatible, and consistent with pertinent standards and laws, data handling can assist organizations in facilitating data interchange and integration across many platforms, systems, or domains. To facilitate safe and smooth data exchange among several organizations in a zero-trust environment, NIST suggests implementing a federated identity management system (Buckbee, 2021). Similar to this, UC Berkeley shows how classification algorithms may be used to analyze and forecast the group membership of various types of data depending on their features and patterns (Imperva, 2021).

# **Challenges of Data Classification and Handling of Documents**

Data classification and handling of documents pose several challenges for organizations, such as: **(1)** Managing the complexity and variety of data sources and kinds. Large quantities and a wide range of data that are produced, gathered, stored, or processed using various forms, locations, or technologies may be a problem for organizations. Applying accurate and consistent data classification and handling standards throughout the company may be challenging as a result. For instance, Microsoft Service Assurance admits that some data types may be more difficult to categorize than others, such as unstructured or dynamic data that may vary over time or rely on the context (Gluckd & Robmazz, 2023). Varonis cautions that some forms of data, such as dark data that is not utilized or analyzed by the organization, may be buried or ignored (Cocchi et al., 2018), **(2)** Aligning diverse stakeholders' business demands and security requirements. Organizations may have to strike a balance between the competing interests or expectations of multiple stakeholders engaged in data generation, ownership, utilization, or governance. Effective communication and coordination among corporate executives, security teams, end users, regulators, customers, partners, or vendors may be required. Gartner recommends incorporating both company owners and security specialists in developing the data categorization framework and levels based on the outcomes of the business impact analysis (BIA) and risk assessment (RA) (Scarfone et al., 2021). Similarly, Imperva suggests forming a cross-functional team to manage the data categorization process and guarantee its alignment with organizational goals (Chugh et al., 2021) and **(3)** Changing the regulatory landscape and dangerous environment. Organizations may be required to comply with multiple or changing standards or regulations governing the protection or disclosure of specific categories of data in various jurisdictions or industries. They may also have to deal with more sophisticated cyberattacks on their sensitive or valuable data assets. This could necessitate ongoing monitoring and modification of their data classification and handling policies, methods, and technologies. To verify the efficacy and compliance of data classification and handling practices, NIST recommends conducting periodic reviews and audits (Buckbee, 2021). Similarly, UC Berkeley proposes comparing the accuracy and effectiveness of multiple systems of classification for different types of data and jobs (Imperva, 2021).

# **Best Practices for Data Classification and Handling of Documents**

Based on the literature review, some best practices for data classification and handling of documents are: **(1)** Establishing a precise and thorough framework for data classification that identifies the classification levels, standards, and labels for various data kinds according to their importance or sensitivity. The framework should be communicated to the entire organization and put into effect. It should also be in line with the organizational goals, risk tolerance, and compliance requirements**, (2)** Putting in place a reliable and expandable data-handling system that, depending on the classification of the data, applies the relevant rules, controls, and actions. As much as practicable, the system should be automated and auditable, and it should relate to the current data platforms, applications, and workflows, and **(3)** Encourage and enable end users to classify and handle their data properly and ethically. The advantages, difficulties, and obligations of data classification and handling should be made clear to end users, and they should be given the direction, resources, and assistance they need to complete their activities efficiently and safely.

**Data Classification Frameworks and Models**

Data classification frameworks and models are tools that help organizations define, communicate, and implement data classification and handling policies. Data classification frameworks provide a set of levels, criteria, and labels for categorizing data based on its sensitivity or impact, while data classification models provide a set of methods, techniques, and standards for applying data classification and handling rules. This compares different types of data classification frameworks and models, highlighting their advantages, disadvantages, and use cases.

Identifying content, classifying it, and protecting it based on its sensitivity or impact are all parts of the data classification process. Data categorization seeks to shield information against unauthorized erasure, modification, or disclosure. A data classification framework is a formal policy typically executed enterprise wide. It often consists of three to five classification levels, which include three elements—name, description, and real-world examples (Tuby & Leshem, 2021). A succinct and descriptive phrase, such as Public, Internal, Confidential, or Highly Confidential, is used in the name to signify how sensitive the data is. The description is a succinct and precise explanation of the significance and range of the data classification level, including any potential repercussions of unauthorized disclosure, change, or destruction.

# **Government classification frameworks**

These are the frameworks that the government's agencies or authorities have prescribed or advised to be used by contractors or organizations in the public sector. They offer a standardized and uniform method of processing data that is subject to legal or regulatory requirements or concerns about national security. According to the likelihood that the material may compromise national security, the U.S. government, for instance, has a four-level classification system: Top Secret, Secret, Confidential, and Unclassified (Gluckd & Robmazz, 2023). Similar to this, the European Union maintains three classification levels for material based on how it can affect the interests of the EU or its member states: EU Top Secret, EU Secret, and EU Confidential (Scarfone et al., 2021).

# **Commercial classification frameworks**

For their own internal usage or external collaboration, these frameworks were created or implemented by private sector organizations. They offer a versatile and adaptable technique to handle data classification and handling that is pertinent to the business goals, risk tolerance, and industry sector of the organization. Like this, Imperva uses a three-level classification system called High-Risk Data (HRD), Medium Risk Data (MRD), and Low-Risk Data (LRD) to categorize data depending on its potential worth to attackers.

# **Data Classification Models**

Data classification models are methods or techniques that help organizations apply data classification frameworks to their data assets (Imperva, 2021). Data classification models can be based on different dimensions or factors, such as content, context, or user input.

* **Content-based classification:** This model categorizes data based on the existence of sensitive information such as personal information, financial information, or intellectual property after analyzing the data's content (Imperva, 2021).
* **Context-based classification:** This model allocates a category depending on the context of the data, such as the source, location, owner, or purpose of the data. For instance, the level of sensitivity of data created by one department, application, or user may differ from that of data created by others (De Groot, 2023).
* **User-based classification:** This model relies on the user's manual input when they create or manage the data and decide which category to put it in. The user has the option to build their custom categories or choose one from a predefined selection. (Imperva, 2021)

Data Classification and Governance

Data classification and governance are critical aspects of managing sensitive information in an organization. Data classification involves categorizing data based on levels of sensitivity and assigning controls to each category (Solove & Hartzog, 2022). Standard data classification levels include public, internal, confidential, and restricted. Public data has minimal sensitivity, while restricted data is highly sensitive (NIST, 2022).

Once data has been classified, governance policies and procedures can be implemented to appropriately manage each category (Kooper et al., 2011). Ladley (2012) defines data governance as "the overall management of data availability, integrity, security, and compliance through standardized processes and decision rights." Developing classification guidelines, implementing controls such as encryption, training employees on handling practices, and performing audits are all examples of key governance activities (Smallwood, 2014).

Access to confidential data may be restricted to specific staff members and require strong encryption (Microsoft, 2022). According to regulations (HIPAA, 1996; PCI DSS, 2022) for restricted data, these measures may be necessary: multi-factor access controls, logging, and encryption both in transit and at rest. Overseeing governance programs are positions in data governance such as Data Protection Officers (Tankard, 2016). Respecting data governance and classification policies lowers risk exposure from theft or unauthorized access.

Understanding what makes data sensitive in the first place is a critical consideration in data classification. According to Solove and Hartzog (2022), sensitivity is determined by the potential harm from unauthorized access or disclosure. Because of the potential consequences of misuse, such as reputational damage or identity theft, data such as intellectual property, customer information, and employee records must be protected (UK ICO, 2020). Data classification based on content, context, and potential impact allows for the application of appropriate governance controls.

Implementing data governance programs necessitates cross-departmental stakeholder participation (Kooper et al., 2011). Policy development, training, auditing, and enforcement should be done collaboratively by legal, compliance, IT, and business units (Ladley, 2012). Data governance frameworks and matrices are used to formalize the distribution of data management responsibilities and decision-making authority (Smallwood, 2014). Ongoing governance also necessitates program adaptation to address new regulations, technologies such as cloud computing, and changing data landscapes (Tankard, 2016). Continuous coordination is required for effective governance.

# **METHODOLOGY**

A mixed-methods methodology would be adopted for this study. Mixed methods research integrates quantitative and qualitative research elements to address the research topic (George, 2023). Mixed methods research gives a more complete picture than isolated quantitative or qualitative studies. It combines the advantages of qualitative and quantitative data collection and analysis methods.

## **Research Design**

The research design would consist of three stages: a literature survey, the construction of a framework, and the evaluation of case studies.

* **Literature review:** To be peer-reviewed, a systematic and comprehensive search of relevant academic databases, journals, books, and papers to discover and synthesize the existing literature on data classification and handling. To choose and analyze sources, the literature review will employ a preset set of keywords, inclusion and exclusion criteria, and quality rating methods. A thematic analysis approach will also be used in the literature review to identify the primary themes, concepts, theories, models, and gaps.
* **Framework development:** a conceptual framework for data classification and processing that combines deductive and inductive reasoning, based on the results of the literature analysis and the goals of the study would be used. The inputs, activities, outputs, consequences, and impacts of data classification and processing will be mapped out by the framework using a logic model. To identify and rank the important players, interests, roles, and duties involved in data classification and management, it will also apply stakeholder analysis.
* **Case study evaluation:** A qualitative and quantitative method for testing and validating the framework in a real-world situation. The case study evaluation will employ a multiple-case design to select and compare two or more cases of data classification and handling in different contexts, a mixed-methods data collection strategy that includes document analysis, interviews, surveys, observations, and experiments, and a mixed-methods data analysis strategy that includes content analysis, descriptive statistics, inferential statistics, and machine learning algorithms.

## **Data Collection**

Data collecting techniques could be document analysis, interviewing, and surveys:

* **Document analysis:** A technique for examining and understanding written or visual documents relevant to the research issue. Document analysis can reveal historical, contextual, or information as well as insights into document makers' or readers' viewpoints, values, and beliefs. Policies, reports, manuals, contracts, emails, websites, social media posts, and other documents can all be analyzed.
* **Interviews:** The technique of having in-person or online discussions with people or groups who have information or experience relating to the study issue. In-depth information can be gleaned from interviews, as well as the interviewees' opinions, attitudes, feelings, and motivations. Depending on the degree of standardization and flexibility sought, interviews might be organized, semi-structured, or unstructured.
* **Surveys:** A technique for gathering data that can be quantified and is standardized from many respondents who represent a population of interest. Surveys can track the frequency, distribution, or connections between variables related to the study's subject. Surveys can be distributed in a variety of ways, including on paper, online, over the phone, or in person.

## **Data Analysis**

The methods for analyzing the data could be machine learning algorithms, descriptive statistics, and content analysis:

* **Content Analysis:** Analyzing the content and interpretation of textual or visual material, such as documents, interviews, or observations. Content analysis can identify and quantify the frequency, patterns, or themes in data, as well as the underlying concepts, classifications, or codes. Content analysis can be deductive or inductive, depending on whether the study is driven by a preset framework or an emergent one.
* **Descriptive Statistics:** Evaluating numerical data from surveys or experiments. Descriptive statistics can summarize and display basic data features such as the mean, median, mode, standard deviation, range, frequency, or distribution. Descriptive statistics can also utilize graphical approaches to visualize data, such as charts, graphs, tables, or histograms.
* **Machine Learning Algorithms:** A method of analyzing numerical or textual data, such as surveys, experiments, or papers. Machine learning algorithms could gain information from data and do tasks that are difficult or impossible for humans to complete manually. Machine learning algorithms can also utilize supervised learning, unsupervised learning, reinforcement learning, classification, clustering, regression, or neural networks to perform tasks like prediction, recommendation, classification, segmentation, or generation.

SUMMARY

In the digital era, the classification of data is essential. When it comes to protecting the security, privacy, and compliance of sensitive data across a variety of sectors, data classification and handling are crucial activities. However, classifying and managing data manually can be time-consuming, unreliable, and prone to mistakes.

The project incorporates a comprehensive approach that includes a clear understanding of data assets, the development of data classification policies and handling procedures, effective employee training, and frequent assessments to guarantee that the system stays effective. The project's research strategy adopts a mixed-approaches approach, combining qualitative and quantitative research methods such as expert interviews, case studies, surveys, focus groups, and document analysis.

Data classification studies the most effective ways to automate and use human skills in data classification and document handling operations.  To create data categorization and document handling systems that make use of automation technology and human expertise, as well as evaluate their effectiveness and usability, government classification and commercial classification frameworks might be offered.

Document management refers to the practice of managing documents during their lifetime, whereas data classification refers to the process of classifying data based on its sensitivity and effect degree. Organizations may secure their data, adhere to regulations, and use their resources more efficiently by using data classification and document management. The complexity and diversity of cloud and dispersed systems, the evolving dangers, and hazards of data breaches, as well as the growing volume and variety of data, present many challenges for data classification and document processing.

Data classification analyses ways to optimize data classification and document handling processes through the use of human skills and automated technology.  Government classification and commercial classification frameworks can be provided for implementing and evaluating data classification and document handling systems that leverage human expertise and automation technologies.

Data classification has benefits such as improving data security and compliance by lowering the risk of data breaches, inadvertent disclosures, or unauthorized modifications, as well as challenges such as managing the complexity and diversity of data sources and types and implementing a robust and scalable data handling system that applies the appropriate rules, controls, and actions for different types of data based on their classification as best practices for data classification.

REFERENCES

# Buckbee, M. (2021, March 25). What is Data Classification? Guidelines and Process. Www.varonis.com. https://www.varonis.com/blog/data-classification

# Chugh, R., Willemsen, B., & Henein, N. (2021, March 31). Building Effective Data Classification and Handling Documents. Gartner. https://www.gartner.com/en/documents/4000054

# Cocchi, M., Biancolillo, A., & Marini, F. (2018). Chemometric Methods for Classification and Feature Selection. 82, 265–299. https://doi.org/10.1016/bs.coac.2018.08.006

# De Groot, J. (2023, March 5). What is Data Classification? A Data Classification Definition. Digital Guardian; Juliana De Groot. https://www.digitalguardian.com/blog/what-data-classification-data-classification-definition

# George, T. (2023). Mixed Methods Research | Definition, Guide & Examples. Scribbr. https://www.scribbr.com/methodology/mixed-methods-research/

# gluckd, & robmazz. (2023, March 3). Data classification & sensitivity label taxonomy - Microsoft Service Assurance. Learn.microsoft.com. https://learn.microsoft.com/en-us/compliance/assurance/assurance-data-classification-and-labels

# Gluckd. (2023, March 2). Data classification & sensitivity label taxonomy – Microsoft Service Assurance. Microsoft Learn. https://learn.microsoft.com/en-us/compliance/assurance/assurance-data-classification-and-labels

# HIPAA. (1996). The Health Insurance Portability and Accountability Act of 1996 (HIPAA). https://www.cdc.gov/phlp/publications/topic/hipaa.html

# Imperva. (2021). What is data classification? | best practices & data types | imperva. Learning Center. https://www.imperva.com/learn/data-security/data-classification/

# Kooper, M.N., Maes, R., & Lindgreen, E.E. (2011). On the governance of information: Introducing a new concept of governance to support the management of information. International Journal of Information Management, 31(3), 195-200. https://doi.org/10.1016/j.ijinfomgt.2010.05.009

# Ladley, J. (2012). Data governance: How to design, deploy and sustain an effective data governance program. Newnes.

# Microsoft. (2022). Data classification for Microsoft 365. https://docs.microsoft.com/en-us/microsoft-365/compliance/data-classification-overview?view=o365-worldwide

# NIST. (2022). Data classification. https://csrc.nist.gov/glossary/term/data\_classification

# PCI DSS. (2022). Payment Card Industry Data Security Standard (PCI DSS). https://www.pcisecuritystandards.org/

# Satori Cyber. (2023, April 22). Data Classification: Compliance, Concepts, and 4 Best Practices - Satori. Satori. https://satoricyber.com/data-classification/data-classification/

# Scarfone, K., Cybersecurity, S., & Souppaya, M. (2021). DATA CLASSIFICATION PRACTICES Facilitating Data-Centric Security Management PROJECT DESCRIPTION. https://www.nccoe.nist.gov/sites/default/files/legacy-files/data-classification-project-description-draft.pdf

# Smallwood, R.F. (2014). Information governance: Concepts, strategies and best practices. Wiley.

# Solove, D.J., & Hartzog, W. (2022). The case for data classification. Yale Journal on Regulation, 39(2). https://digitalcommons.law.yale.edu/yjreg/vol39/iss2/5

# Tankard, C. (2016). What does the GDPR mean for business. Network Security, 2016(6), 5-8. https://doi.org/10.1016/S1353-4858(16)30056-3

# Tuby, S., & Leshem, E. (2021, December 15). Data Classification Framework: What, Why, and How. Satori. https://satoricyber.com/data-classification/data-classification-framework-what-why-and-how/

# UK ICO. (2020). What makes personal data sensitive? https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/what-makes-personal-data-sensitive/

# Web service, amazon. (n.d.). Data classification models and schemes - Data Classification. Docs.aws.amazon.com. Retrieved May 5, 2023, from https://docs.aws.amazon.com/whitepapers/latest/data-classification/data-classification-models-and-schemes.html

# 